The NCAR Community Data Portal (CDP)
Experiences with OAI metadata record federation

http://cdp.ucar.edu/

presented by

Michael Burek

(NCAR/SCD/VETS)

Acknowledgments:

CDP staff: Dave Brown, Luca Cinquini, Don Middleton (PI), Markus Stobbs, James Humphrey

funding: NCAR's directorate, NSF

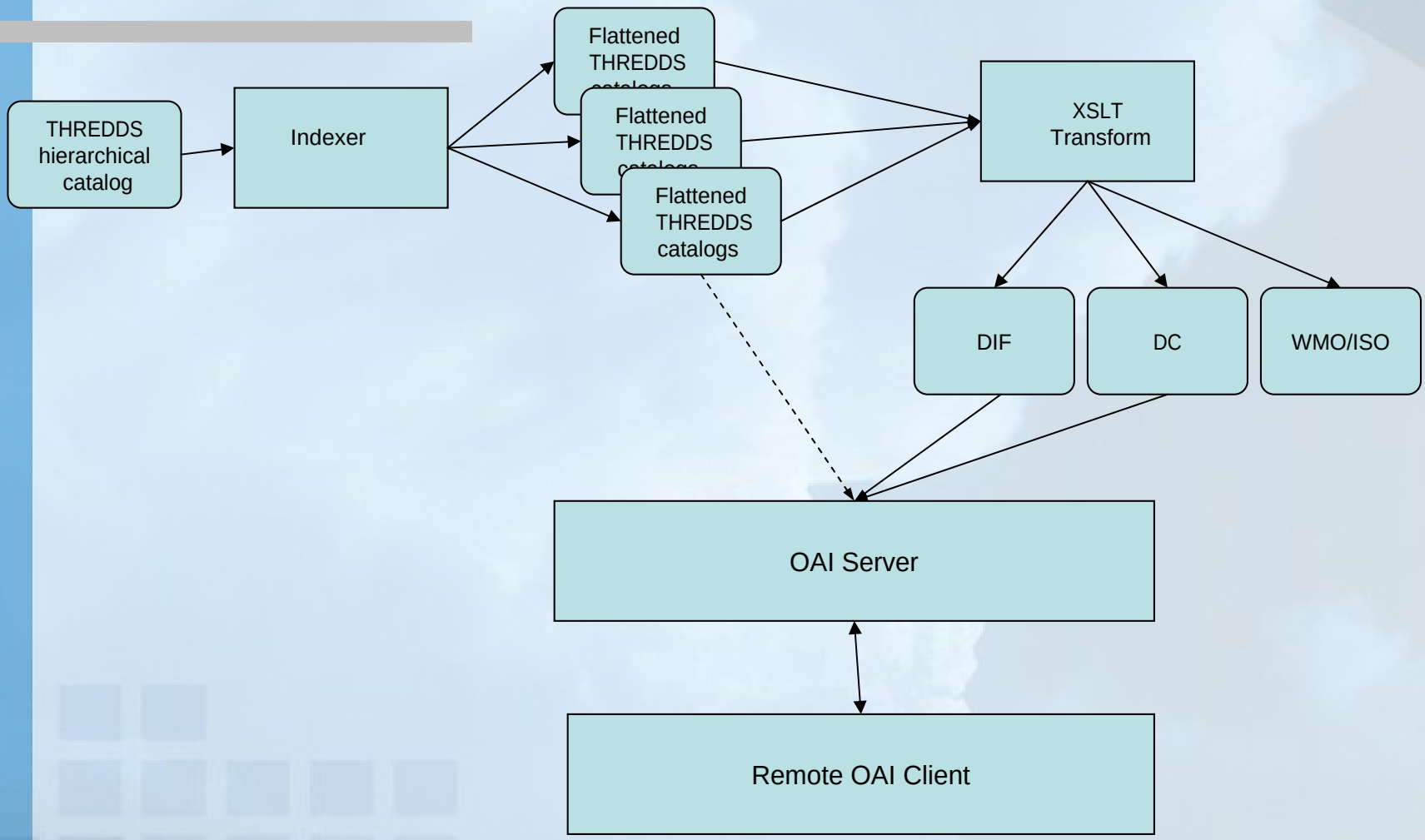
Introduction, What is OAI

- OAI: Open Archives Initiative
- Goal: Provide a lightweight infrastructure for sharing metadata records among participating institutions
- OAI Began in 199x to serve the library and e-print communities
- OAI model consists of six verbs -- Identify, List Metadata Formats, List Sets, List Identifiers , List Records, Get Records
- OAI base mode specifies Dublin Core as the default schema for shared records but makes provision for other schemas to be used

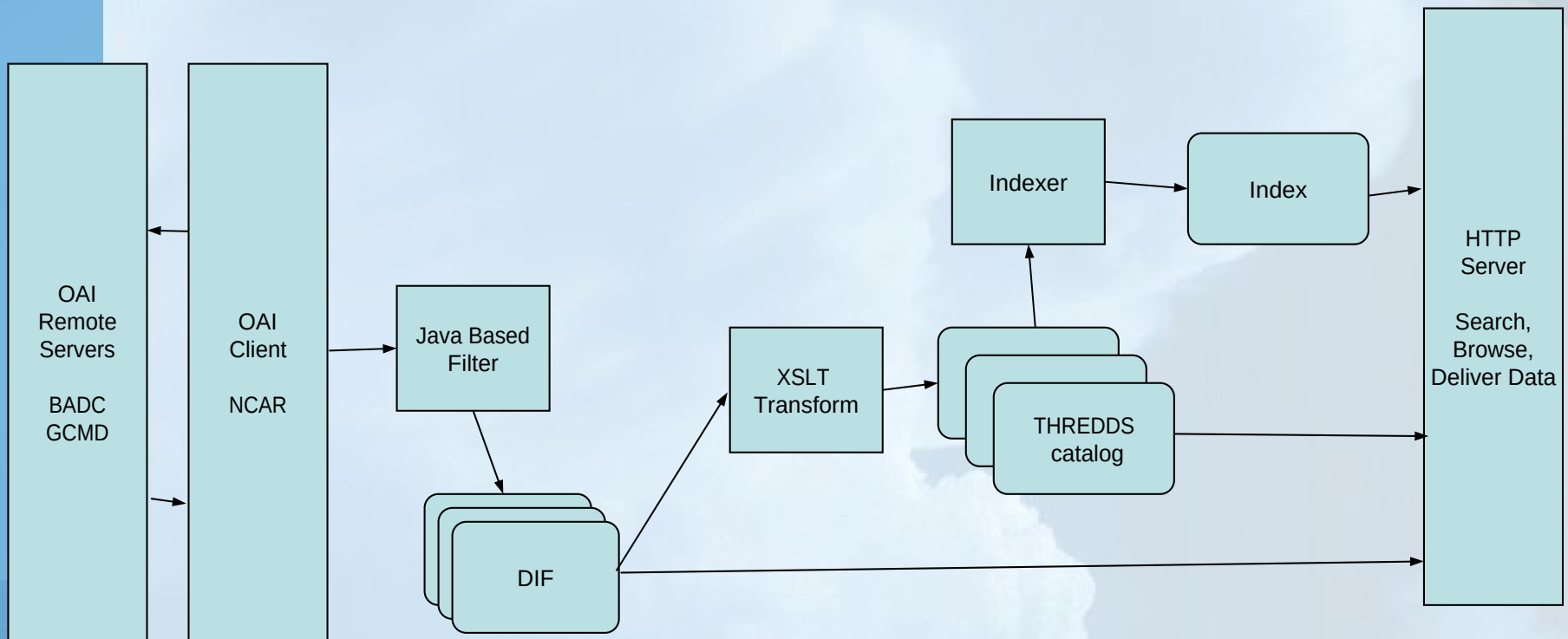
OAI record sharing effort NCAR, BADC, GCMD

- GCMD DIF records were shared
- INCOMING Records:
  - DIF records harvested by NCAR were transformed into THREDDS schema using XSLT
  - The transformed THREDDS records were ingested into the CDP Search and Browse functions
  - Links to BADC data that were included in the generated DIF records enabled linking back to BADC data from the CDP search and browse
- OUTGOING Records:
  - CDP hierarchical THREDDS catalogs were "flattened" into transformable THREDDS catalogs then, into DIF and DC
  - NCAR DIF records were harvested by BADC and GCMD
  - NCAR DC records were harvested by the University of Michigan Digital Library Oaister project
  - DC records were sent to the Yahoo search engine via UM

Current CDP OAI server architecture



OAI metadata Harvest



Demo

- BADC records on CDP
- Search on oaister site
- Demo finding NCAR records on BADC site

OAI Harvesting Technical Issues

- Many extra items in the records that cause problems:
  - Records are marked as UTF-8 but in fact are ISO 8859-x
    - UTF-8 and ISO 8859-x are incompatible!
    - UTF-8 and 7 bit ASCII *are* compatible, which leads to confusion
  - Embedded HTML in text fields(using escaped <> symbols, or not)
    - Usually a byproduct of automatic creation of legacy records
  - Text fields contain un-escaped special characters (&, <, >, /)
    - Even if allowed in one schema it may not transform well to another e.g. a text field that transforms to an element attribute in another.
  - Namespace issues
    - OAI requires a schema
    - GCMD is still in the process of creating a schema for DIF
    - NCAR and BADC created working DIF schemas in the meantime, not quite compatible

CONCLUSION: There needs to be filtering code before the XSLT transform

OAI schema / XSLT transformation issues

- Schemas do not always agree in level of detail
- Schemas have different required elements
- Schemas can have different controlled vocabularies
- Adoption of dataset identifiers that don't overlap is crucial
  - All centers currently have their own standards
- OAI does not address duplication of records, need to establish republication "rules of the road" to avoid problems

OAI future directions

- GO-ESSP could consider creating/adopting a community wide OAI metadata interchange schema that addresses the issues on earlier slide, (similar to DC, or may be an existing variant of DC)
- RECOMMENDATAION: GO-ESSP coordinate community wide record ID schema (adopt DC's methods?)
- Collaborations with other GO-ESSP institutions
- GO-ESSP could provide some tutorial support.
  - XML do's and don'ts
  - Identifier guidance
  - Rules of the road for republication

Questions?